Rental price estimation

Oskar Grøtting Strand, 19.10.2025

1: DESCRIBE THE PROBLEM

SCOPE

Project Goal:

The goal of this project is to build a machine learning based rental price prediction system for Norway. Given user input such as area, rooms and size of the apartment, the system predicts monthly rent for the year 2025. The solution uses historical rental data from SSB from the years 2012 to 2024 and applies a regression model to try and estimate a future price.

Rental prices are influenced by multiple factors such as location, size, and type of housing. This system aims to help users figure out what rental prices they can expect when moving to a new city, area, upsizing or downsizing their current apartments. Intended users will mostly be students and young adult which do not have the economy to buy and are looking to rent. When using the system, the user simply inputs desired location, size in square meters and number of rooms.

There exist some solutions today, one for example is leiepris.husleie.no, but mainly people compare different listing on sites such as finn.no and hybel.no and figure out from there which places are affordable for them.

A manual solution would be solved by averaging past prices from each area and just applying a flat growth percentage, however, this would be time consuming if there is no public dataset. And in the future if you wished to scale the data for a better prediction later, it would be smarter to have a ML model, as they work even better with larger datasets. As a business impact this would help users, know their local market prices, avoid overpaying rent, have a transparent renting market. This is something everyone wants and will ensure the product succeeds in the market.

METRICS

The model's performance will be evaluated using three complementary metrics that measure prediction accuracy:

Mean Absolute Error (MAE) will be the primary metric, as it represents the average prediction error in kroner and is easy to interpret. The target is MAE below 1,500 kr, meaning predictions should typically be within 10-15% of actual prices for most apartments. This level of accuracy is sufficient for users to budget and plan.

R^2 (Coefficient of Determination) will measure how much of the price variation the model can explain. The target is R^2 above 0.80, meaning the model should capture at least 80% of the factors that determine rental prices.

Root Mean Squared Error (RMSE) will be used to detect large prediction errors, as it penalizes bigger mistakes more heavily than MAE. The target is RMSE below 2,000 kr to ensure the model doesn't make catastrophic mispredictions that could mislead users.

Software Performance Metrics:

Beyond prediction accuracy, the system's usability will be measured through:

Response latency - The time from user input to displayed prediction should be under 1 second for 95% of requests.

User interface validation - Dropdown menus will only show valid combinations (preventing queries for impossible apartment sizes)

Business Objectives:

These metrics directly support the business goal of providing quick, reliable rental price estimates for planning purposes. MAE below 1,500 kr ensures users can budget with reasonable confidence. R^2 above 0.80 builds trust by showing the model captures real pricing patterns. Response time under 1 second makes the tool practical for real-time use, increasing the likelihood users will complete their searches and continue to use the service.

Success Criteria:

The project will be considered successful if:

- **Technical performance:** MAE $\leq$ 1,500 kr and $R^2 \geq 0.80$ on validation data.
- **Usability:** Predictions delivered within 1 second and intuitive interface preventing invalid queries.
- **Practical value:** Estimates within ±15% of actual prices for at least 75% of queries, making them useful for renters budgeting, landlords setting rates, and real estate professionals doing market checks.

2: DATA

The data used in this project consists of historical rental prices in Norway between 2012 and 2024 collected from Statistisk Sentralbyrå (SSB), each data point contains:
Features: Geographical area (Område), year(år), and number of rooms(rom) with its size.
Target variable: rental prices monthly in NOK

Since the dataset was retrieved from SSB's official website the data is highly reliable, and consistent as it comes from a trusted government source

The dataset contains 401 rows with 15 columns, during the project we identified two issues.

Small dataset size: in the dataset there was less data than originally expected when it was retrieved.
Uneven distribution: big cities like Oslo, and Trondheim had more data from different urban areas than smaller cities.

Before settling on rental price data, several alternatives were explored:

- **Fantasy Premier League data:** Too large and complex for project scope
- **Yahoo Finance API (yfinance):** Required extensive cleaning and feature engineering

The SSB rental data was chosen for its balance of cleanliness, reliability, and appropriate scope for this project.

Important notes about the data from SSB:

- **Survey nature**: LMU is primarily a level measurement survey where comparisons over time should be made with caution
- **2019-2020 discontinuity**: Due to extensive municipal reform and changes in how centrality is handled in SSB's regression model, data from 2020 onwards is less comparable to earlier years
- **Price methodology**: Prices are predicted values from SSB's regression model, not raw transaction data
- **Included**: Rental contracts with professional rental companies and private landlords (excluding family/friends arrangements)
- **Excluded**: Electricity/heating costs, garage fees, balcony/terrace premiums, and furniture premiums
- **Uncertainty**: Prices are rounded to the nearest hundred kroner from 2013 onwards due to estimation uncertainty, which is highest for very small and very large units
- **Regional definitions**:
    - "Akershus - nærliggende Oslo kommuner" includes: Ski, Nesodden, Oppegård, Bærum, Asker, Lørenskog, and Skedsmo
    - "Utkant Akershus" includes remaining municipalities
- **Apartment size limits**: Units with more than 8 rooms are excluded from calculations

The dataset does not contain any personal information or sensitive attributes. It is fully anonymized by SSB and therefore there is no GDPR concerns.

The preprocessing steps were as following:
The raw SSB data was transformed into a long-format for model training. The following transformations were applied.

- One-hot encoding on categorical columns such as “Område”.
- Feature extraction, “Størrelse” column was split into “rom” and “størrelse_m2”.
- Rows with missing “leiepris” were removed in the case they existed.
- 2024 was held out as the validation year.

- Region specific growth was added for the 2025 predictions.

Despite having less data than initially planned, the dataset is still sufficient for building a regression model.

End of project realisation:
A significant limitation of this project is that the data consists of predicted values from SSB's own regression model rather than actual observed rental prices. This means our model is essentially "learning from another model's predictions" rather than real-world data. Ideally, raw data from rental contracts would be used, but such data is not publicly available.

3: MODELING

A baseline performance value could be established by taking average rental prices of an area and applying a simple inflation-based price increase based on the years calculated inflation, to get an estimated price index for the different regions. Although this is not a machine learning solution it is a useful baseline to compare future models with.

When modeling, research indicated that regression models would be most suitable for the continuous target variable. The choice was between Ridge Regression and XGBoost. XGBoost was tested first, as it performs well on structured datasets and can capture complex patterns. However, after training a basic model with default parameters, the results appeared too good to be true. Analysis of the learning curve revealed severe overfitting the model achieved near-perfect accuracy on training data but poor performance on validation data. This indicated the model was memorizing the training examples rather than learning generalizable patterns, making it unsuitable for the small dataset of 401 rows.

Ridge Regression was selected as the final model because its L2 regularization is well-suited for small datasets and helps prevent overfitting. The model was trained on 2012-2023 data with hyperparameter tuning performed using GridSearchCV and TimeSeriesSplit.

Analysis of prediction errors revealed bias the model underestimated prices in expensive urban areas while overestimating in smaller towns. This indicates the linear model struggles to capture non-linear pricing relationships in high-end locations.

For 2025 predictions, a two-step approach was implemented. Ridge predicts the 2024 base price, then area-specific growth rates calculated from 2023 to 2024 trends adjust the estimate forward to 2025.

4: DEPLOYMENT

The model is deployed as a web application using streamlit. Users can then access the site and through drop down menus select area, rooms and apartment size. The app then returns an

estimate price for 2025 based on the inputs. They way to improve the systems after deployment would be to retrain the model annually as new rental data becomes available. The model as of now works well in medium sized cities such as Tromsø, Stavanger and Kristiansand. Other than that, it could be possible to get user feedback on the estimations for future improvement.

5: REFERENCES

Data:

- Statistics Norway (SSB). (2025). «Leie av bolig, månedlig leie (LMU)»
  Retrieved from: https://www.ssb.no/statbank/table/09897/
  Accessed: 21/10/2025

Other:

- Høgskulen på Vestlandet. (2025). DAT158 Maskinlæring – Lecture notes and syllabus. [Autumn/2025]
- OpenAI. (2025). ChatGPT(Version 5). Used for: Coding assistance and feedback
- Claude. (2025) Claude(Sonnet 4.5). Used for: Evaluation on finalising product
- StatQuest with Josh Starmer - XGBoost in Python from Start to Finish (YouTube) https://www.youtube.com/watch?v=GrJP9FLV3FE
- StatQuest with Josh Starmer. *Ridge Regression: Clearly Explained!!!*. YouTube. https://www.youtube.com/watch?v=Q81RR3yKn30
- Geeksforgeeks. Regression metrics. https://www.geeksforgeeks.org/machine-learning/regression-metrics/